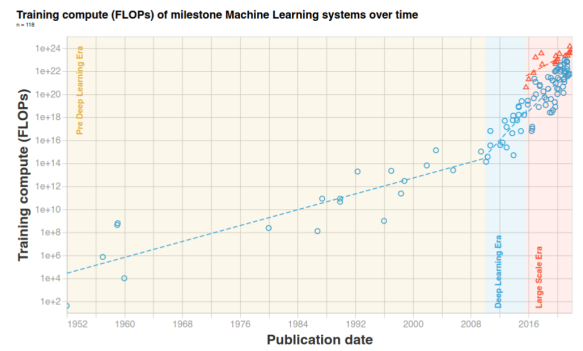
L'Éfficacité Réduit-elle les Impacts Environnementaux des Entraînements de Modèles d'Apprentissage Machine ? Une perspective temporelle

Clément Morand, Anne-Laure Ligozat & Aurélie Névéol

25 mars 2025

Université Paris-Saclay, Laboratoire Interdisciplinaire des Sciences du Numérique, CNRS

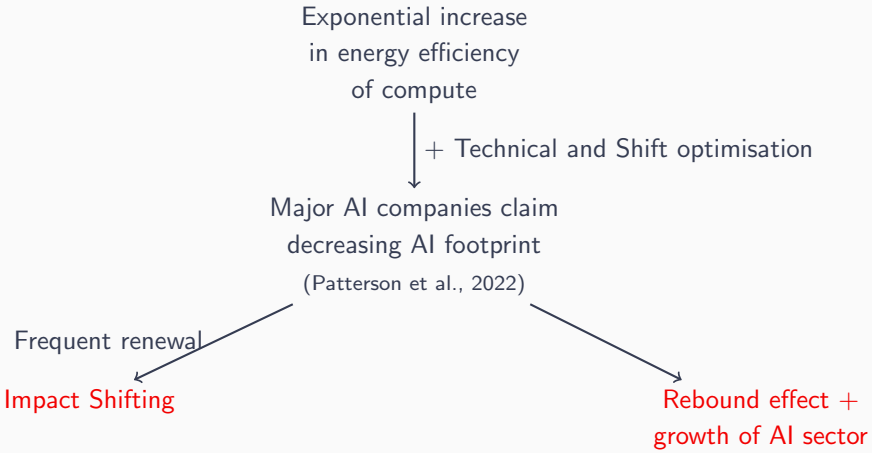
Machine Learning requires an ever-increasing amount of compute to train



(Sevilla et al., 2022)

- 10 - 15 % of Google's energy consumption (Patterson et al., 2022)
- Important emissions from energy consumption : 552 tCO_2e to train GPT-3 once and 38 tCO_2e for BLOOM (Luccioni et al., 2023)

Numerous optimisations: how are the impacts of compute evolving?



Methodology

Gathering information on Graphics cards for Machine Learning

TechPowerUp GPU database

Wikipedia list of NVIDIA graphics cards

Other sources (e.g., Google documentation)

NVIDIA Workstation graphics cards between 2013 & 2023

specific cards (e.g., Google's TPU)

- 173 cards models
- 83 cross-validated (47%)
- NVIDIA datasheets when diverging

- *Thermal Design Power* (TDP)
- GPU die area and technological node
- memory type and size
- compute power
- release date

Data on Machine learning training

EpochAI Notable ML systems dataset (Epoch AI, 2024)

> *Models that have advanced the state of the art, had a large influence in the field's history, or had a large impact within the world.*[1]

Required information to estimate the environmental damages of model training:

- training duration
- training hardware
- electricity source

[1] https://epochai.org/data/notable-ai-models-documentation

Information on training duration: number of GPU hours

If training duration and number of cards are available

- 131 models (14% of entries)
- *GPU hours* = *training duration* $\times$ $\#$*cards*
- most reliable estimate as it uses information directly from papers presenting models

Information on training duration: number of GPU hours

If training duration and number of cards are available

- 131 models (14% of entries)
- $\textit{GPU hours} = \textit{training duration} \times \#\textit{cards}$
- most reliable estimate as it uses information directly from papers presenting models

If Training hardware and number of FLOP during training are available

- 103 other models ($\sim$ 25% of entries in total)
- $\textit{GPU hours} = \frac{\#\textit{FLOPS}}{\textit{peak performance}}$
- linear regression to predict performance ratio when both estimates are available (100 observations)
- predicts $\sim$ 27% constant performance ratio

Training hardware characteristics

values consistent with hyper-scaler datacenters

- 2 CPU per server plus:
  - NVIDIA workstation cards: 4 graphics cards
  - NVIDIA non-workstation cards: 2 graphics cards
  - non-NVIDIA cards: manufacturer documentation for the number of cards
- 512 GB memory per workstation server, 192 GB per non-workstation server
- 3 year server duration based on graphics card lifespan (Ostrouchov et al., 2020)
- Information from META: average utilization of 50% (Wu et al., 2022)

increased datacenter efficiency from 2012 to 2018
linear interpolation from average datacenter PUE (~ 1.75) in 2010 to hyperscaler PUE (1.2) from 2018 onwards.

Electricity source and modeling carbon intensity optimisation

Use the carbon intensity of the country of the ML system producer
If multiple countries are involved, all are considered to create a value interval

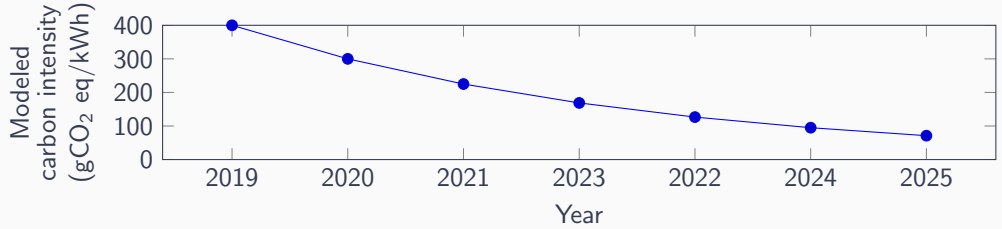
Electricity source and modeling carbon intensity optimisation

Use the carbon intensity of the country of the ML system producer
If multiple countries are involved, all are considered to create a value interval

Modeling strategies for reducing the environmental impact of energy usage

- Aims at accounting for compute location shifting and investment for de-carbonizing data-center electricity sources
- Continuous reduction of the carbon intensity of up to 25% per year starting in 2019.

Example (Modeled evolution of the carbon intensity of the USA electicity mix:)



Assessing environmental damages

Using the MLCA tool (Morand et al., 2024)
Bottom-up approach to evaluate hardware production and usage based on hardware characteristics and information about training process

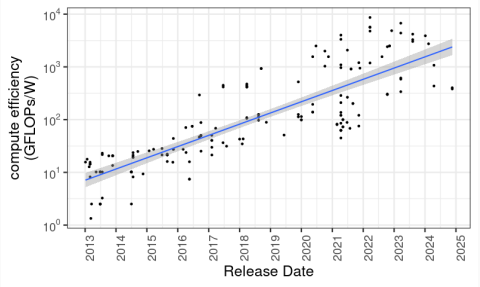
Assesses:

- Carbon footprint through *Global Warming Potential* (GWP100, expressed in $kgCO_2$ eq)
- Metalic resource depletion through *Abiotic Resource Depletion* (ADP, expressed in kgSb eq)

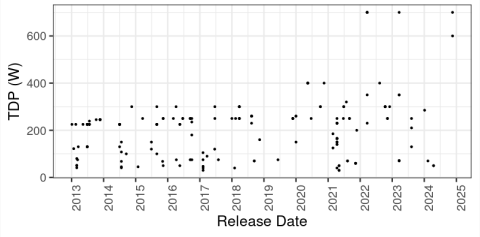
Results

Energy efficiency to scale-up compute

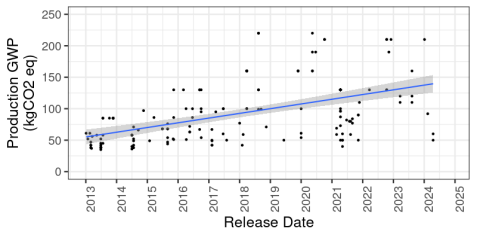
Exponential increase in compute efficiency



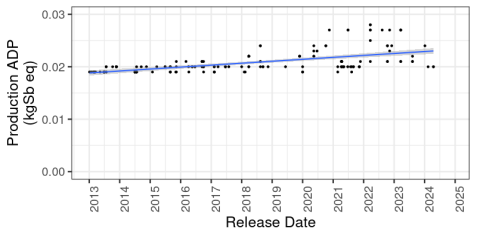
but slightly increasing total energy consumption



Increase in the environmental damages of produced graphics cards

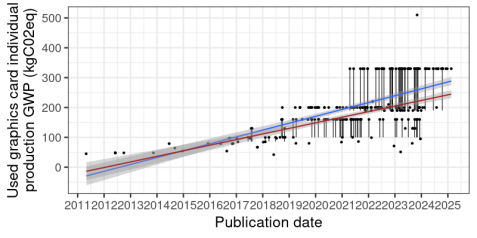


Carbon Footprint

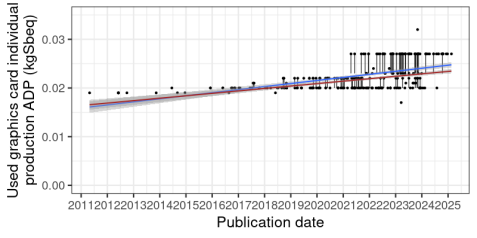


Metallic resource depletion

Increase in the environmental damages of graphics cards used

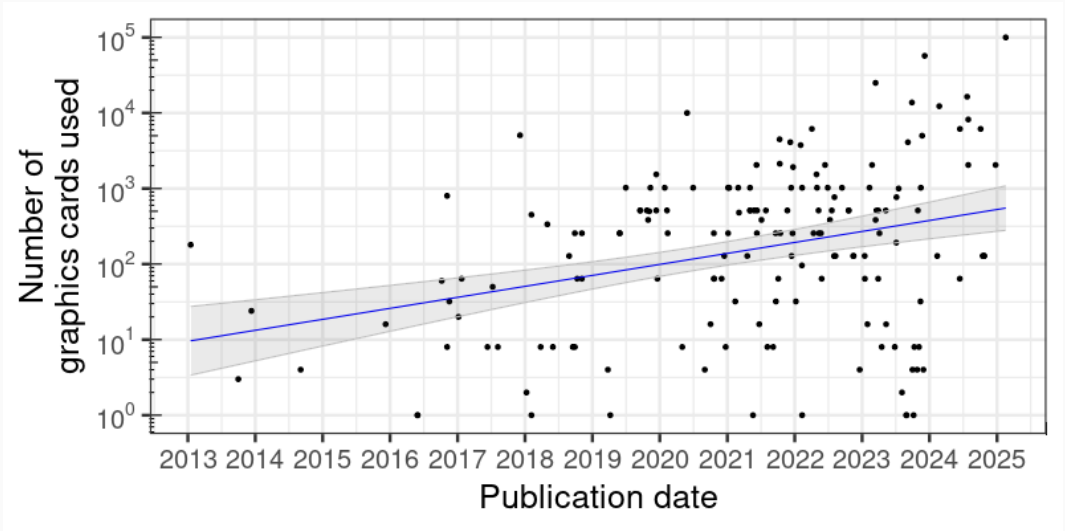


Carbon Footprint

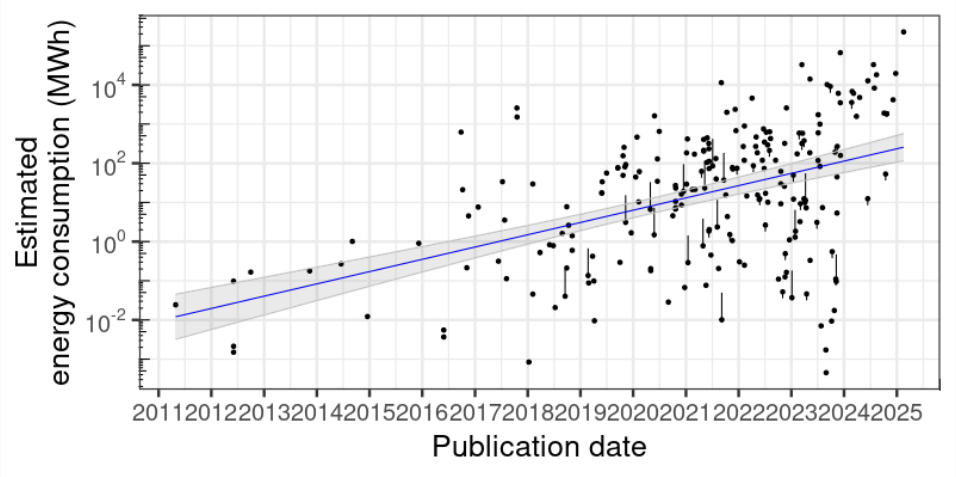


Metallic resource depletion

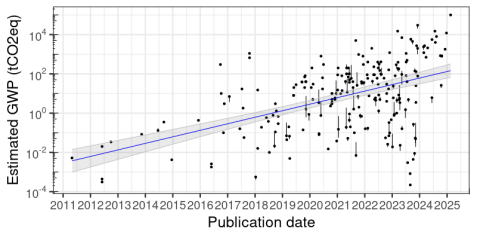
Large increase in the number of cards to train models



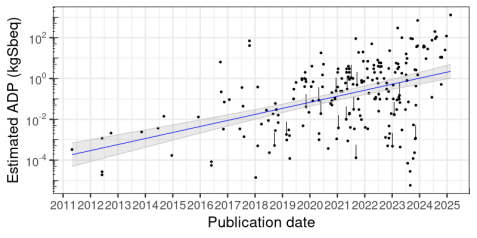
Exponential increase in the energy consumption of models training



Exponential increase in the environmental damages of models training

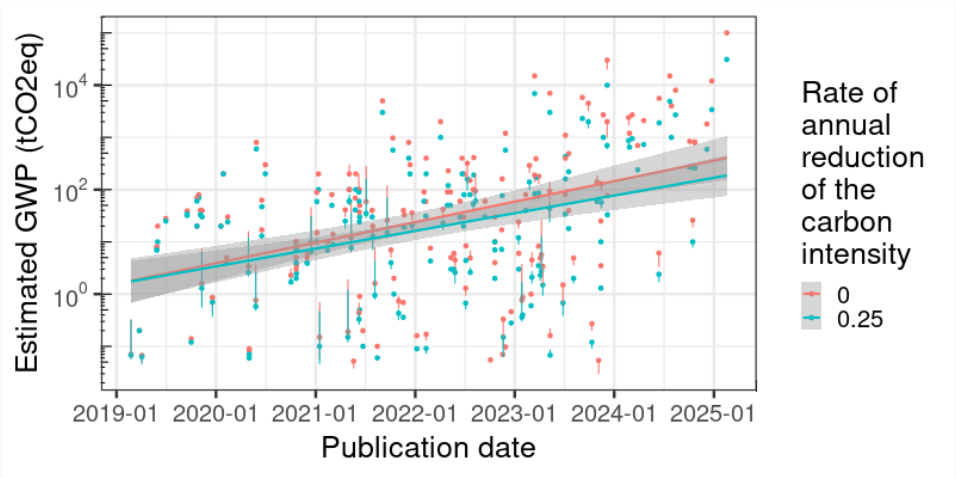


Carbon footprint



Metallic resource depletion

Greener energy cannot void carbon footprint of models training



Conclusion

Current impact reduction strategies alone cannot curb the growth in the environmental impacts of AI training.

- Impacts are partly shifting to the production phase
- Increase in the environmental damages of producing graphics cards
- Optimizations have served scaling-up and not scaling down
- Growth paradigm for machine learning models translates into an exponential growth of the energy consumption and environmental damages of models training
- Need to combine impact reduction strategies with broader reflection on the place and role of AI in a sustainable society.

References

References

- Epoch AI. (2024). Data on notable ai models [Accessed: 2025-02-28]. https://epoch.ai/data/notable-ai-models
- Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2023). Estimating the carbon footprint of BLOOM, a 176b parameter language model. *Journal of Machine Learning Research*, *24*(253), 1–15. http://jmlr.org/papers/v24/23-0069.html
- Morand, C., Ligozat, A.-L., & Névéol, A. (2024). MLCA: a tool for Machine Learning Life Cycle Assessment. *2024 10th International Conference on ICT for Sustainability (ICT4S)*, 227–238. https://doi.org/10.1109/ICT4S64576.2024.00031

References ii

- Ostrouchov, G., Maxwell, D., Ashraf, R. A., Engelmann, C., Shankar, M., & Rogers, J. H. (2020). Gpu lifetimes on titan supercomputer: Survival analysis and reliability. *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–14. https://doi.org/10.1109/SC41405.2020.00045
- Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D. R., Texier, M., & Dean, J. (2022). The carbon footprint of machine learning training will plateau, then shrink. *Computer*, *55*(7), 18–28. https://doi.org/10.1109/MC.2022.3148714
- Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). Compute trends across three eras of machine learning. *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN55064.2022.9891914

References iii

Wu, C., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Behram, F. A., Huang, J., Bai, C., Gschwind, M., Gupta, A., Ott, M., Melnikov, A., Candido, S., Brooks, D., Chauhan, G., Lee, B., Lee, H. S., ... Hazelwood, K. M. (2022). Sustainable AI: environmental implications, challenges and opportunities. In D. Marculescu, Y. Chi, & C. Wu (Eds.), *Proceedings of machine learning and systems 2022, mlsys 2022, santa clara, ca, usa, august 29 - september 1, 2022*. mlsys.org. https://proceedings.mlsys.org/paper/2022/hash/ed3d2c21991e3bef5e069713af9fa6ca-Abstract.html